



SURFACE VEHICLE INFORMATION REPORT

J3329™

APR2026

Issued

2026-04

Artificial Intelligence in Ground Vehicles: Technical, Operational, and Regulatory Challenges

RATIONALE

Given the rapid interest in Artificial Intelligence (AI) and deployment of AI-based solutions, it is important to develop a good understanding of the key challenges relevant to AI technology applications to light and heavy ground vehicles and transportation infrastructure.

As such, a list of technical, operational, and regulatory challenges pertaining to safe, secure, and efficient implementation of AI is documented with recommendations on how to address these challenges based on current state-of-the-art engineering practice and information from industry international standards, technical reports, and technical specifications.

TABLE OF CONTENTS

1.	SCOPE.....	3
2.	REFERENCES.....	3
2.1	AI Systems.....	4
2.2	Model Quality.....	4
2.3	Data Management.....	5
2.4	Data Quality.....	5
2.5	Testing and Assessment.....	6
2.6	Computing Devices.....	6
2.7	Standardization.....	7
2.8	Compatibility Challenges with V-Model.....	7
2.9	Risk Management.....	8
2.10	Trustworthiness.....	9
2.11	Regulatory Challenges.....	9
3.	AI SYSTEMS.....	10
3.1	Main AI Techniques Classifications.....	10
3.2	Difference Between AI and Convnetional Methods.....	11
3.2.1	Inductive versus Deductive.....	11
3.2.2	Data-Driven versus Rule-Based.....	12
3.2.3	End-to-End versus Modular.....	12
4.	TECHNICAL CHALLENGES.....	14
4.1	ML Models Construction.....	14
4.1.1	Different Training Data Split.....	14
4.1.2	Different Training Parameters.....	15
4.1.3	Instance Level Performance.....	15
4.1.4	Model Compression and Pruning.....	16
4.1.5	Model Fine-Tuning and Update.....	16
4.1.6	Model Quality.....	17

SAE Executive Standards Committee Rules provide that: "This report is published by SAE to advance the state of technical and engineering sciences. The use of this report is entirely voluntary, and its applicability and suitability for any particular use, including any patent infringement arising therefrom, is the sole responsibility of the user."

SAE reviews each technical report at least every five years at which time it may be revised, reaffirmed, stabilized, or cancelled. SAE invites your written comments and suggestions.

Copyright © 2026 SAE International

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, transmitted, in any form or by any means, electronic, mechanical, photocopying, recording, or otherwise, or used for text and data mining, AI training, or similar technologies, without the prior written permission of SAE.

TO PLACE A DOCUMENT ORDER: Tel: 1-877-606-7323 (U.S. and Canada only)
Tel: 1-724-776-4970 (outside U.S. and Canada)
Fax: 724-776-0790
Email: CustomerService@sae.org
<http://www.sae.org>

SAE WEB ADDRESS:

Full document details: SAE J3329_202604



Documents

Document Sharing Hub

For more information on this standard, visit

https://www.sae.org/standards/content/J3329_202604/

4.2	Data Management and Quality	18
4.2.1	Data Management.....	18
4.2.2	High Dimensionality	20
4.2.3	Complicated Real-World Data Distribution	21
4.2.4	Operational Environment Change in Model Deployment.....	21
4.2.5	Distributional Consistency Assumption.....	22
4.2.6	Data Quality	23
4.3	Testing and Assessment.....	24
4.3.1	ISO/IEC TR 29119-11	24
4.3.2	Other Standards.....	25
4.4	Computing Devices	27
5.	OPERATIONAL CHALLENGES	29
5.1	Standardization	29
5.2	Compatibility Challenges with V-Model.....	32
5.3	Risk Management	34
5.3.1	Privacy.....	36
5.3.2	Cybersecurity	36
5.3.3	Safety	38
5.4	Engineering Effort	40
5.5	Environmental Impact	41
5.6	Trustworthiness	41
5.6.1	Ethics and Societal Considerations	42
5.6.2	Transparency	42
5.6.3	Bias	43
5.6.4	Controllability.....	45
6.	REGULATORY CHALLENGES	46
6.1	United Nations and OECD	46
6.2	European Union AI Act.....	47
6.2.1	Risk-Based Approach	47
6.2.2	High-Risk AI Systems	48
6.3	The Rest of the World	51
6.4	AV Regulations	51
7.	NOTES.....	52
7.1	Revision Indicator.....	52
Figure 1	Big data application provider sub-roles and activities.....	19
Figure 2	Computing devices hierarchy.....	27
Figure 3	Safety assurance uncertainty versus complexity	30
Figure 4	AI RMF functions.....	35
Figure 5	Encompassing system development versus AI system development	38
Figure 6	AI risk pyramid	48
Table 1	Inductive AI versus deductive conventional methods	11
Table 2	Data-driven AI versus rule-based conventional methods	12
Table 3	End-to-end AI versus modular conventional	13
Table 4	Use cases complexity and safety assurance comparison	31
Table 5	Trustworthiness comparison between AI and traditional systems	42

1. SCOPE

This SAE Information Report identifies and documents the AI implementation challenges in the following areas:

- **Technical Challenges (see Section 4):** Focusing on the technical hurdles to develop AI models from data for complex human-like functions such as recognition, comprehension, and decision-making. Some AI technologies that do not necessarily involve learning from data, such as search algorithms, will not be considered.
- **Operational Challenges (see Section 5):** Focusing on the unique difficulties to deploy AI in ground vehicles and supporting infrastructure. These difficulties arise, for example, from issues like cost, environmental concerns, safety, security, etc.
- **Regulatory Challenges (see Section 6):** AI-related regulations are rapidly evolving. This section provides an overview of the key AI regulations at the present and some of the challenges to meet them in the ground vehicle domain.

Where applicable, this technical report also provides references to AI-related International Standards (IS), Technical Reports (TR), Technical Specifications (TS), and Information Reports (IR) produced by the International Organization for Standardization (ISO), the Institute of Electrical and Electronics Engineers (IEEE), and SAE International (SAE) that address the challenges listed. For brevity, the authors of this report use the general term “Standards” to reference all the above types of documentation referenced. The full list of these standards is included in Section 2. In addition, the authors of this report are restricting the standard’s scope to only published documents and did not include any other type of literature due to space limitations and keeping the report focused.

The functional definitions and noted issues and concerns are provided consistent with current industry mobility practices and published peer-reviewed literature. The mitigation examples in this document are not concepts that are necessarily independent of each other. Additionally, some mitigations can conflict with others. For the development of a project, an approach is to use systems engineering to design and implement mitigations, synergizing where possible and performing trade-offs where needed to achieve the system requirements (including applicable regulations and standards).

2. REFERENCES

Under the general umbrella of “standards,” the authors of this SAE Information Report have narrowed the scope of the reference information investigated for this report to the following publication types: International Standards (IS), Technical Report (TR), Technical Specification (TS), Publicly Available Standards (PAS), and Information Reports (IR). Selected published documents from the following sources were considered in this report:

- International Organization for Standardization (ISO): Selected publications by:
 - ISO/IEC JTC 1/SC 42 (artificial intelligence) - All
 - ISO/IEC JTC 1/SC 27 (information security, cybersecurity, and privacy protection) - AI relevant only
 - ISO/IEC JTC 1/SC 7 (software and systems engineering) - AI relevant only
 - ISO/TC 22/SC 32 (electrical and electronic components and general system aspects); limited to publications by the following Work Groups: WG 8 (Functional Safety), WG 11 (Cybersecurity), WG 13 (Safety for driving automation systems), and WG 14 (Safety and Artificial Intelligence) - All publications after 2017
- Institute of Electrical and Electronics Engineers (IEEE): Selected publications released by the Artificial Intelligence Standards Committee
- SAE International: Information Reports on the topic of AI application in automotive domain.

2.1 AI Systems

These standards, referenced in Section 3, provide basic concepts and terminologies for AI/ML systems and examples of use cases of AI systems.

No.	Publisher	Publication
1	SC 42	ISO/IEC 22989:2022 - Information technology - Artificial intelligence - Artificial intelligence concepts and terminology
2	SC 42	ISO/IEC TR 24372:2021 - Information technology - Artificial intelligence (AI) - Overview of computational approaches for AI systems
3	SC 42	ISO/IEC 23053:2022 - Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)
4	SC 42	ISO/IEC TR 24030:2024 - Information technology - Artificial intelligence (AI) - Use cases
5	SC 42	ISO/IEC 5339:2024 - Information technology - Artificial Intelligence - Guidance for AI applications
6	SC 42	ISO/IEC 5392:2024 - Information technology - Artificial intelligence - Reference architecture of knowledge engineering

2.2 Model Quality

These standards, referenced in 4.1.6, provide descriptions of how the Software Quality Requirements and Evaluation (SQuaRE) methodology applies to AI systems.

No.	Publisher	Publication
7	SC 42	ISO/IEC 25059:2023 - Software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Quality model for AI systems
8	SC 42	ISO/IEC TS 25058:2024 - Systems and software engineering - Systems and software Quality Requirements and Evaluation (SQuaRE) - Guidance for quality evaluation of artificial intelligence (AI) systems

2.3 Data Management

These standards, referenced in 4.2.1, provide basic concepts related to management of Data for AI systems with emphasis on Big Data technology.

No.	Publisher	Publication
9	SC 42	ISO/IEC 20546:2019 - Information technology - Big data - Overview and vocabulary
10	SC 42	ISO/IEC 24668:2022 - Information technology - Artificial intelligence - Process management framework for big data analytics
11	SC 42	ISO/IEC TR 20547-1:2020 - Information technology - Big data reference architecture - Part 1: Framework and application process
12	SC 42	ISO/IEC TR 20547-2:2018 - Information technology - Big data reference architecture - Part 2: Use cases and derived requirements
13	SC 42	ISO/IEC 20547-3:2020 - Information technology - Big data reference architecture - Part 3: Reference architecture
14	SC 27	ISO/IEC TR 20547-4:2020 - Information technology - Big data reference architecture - Part 4: Security and privacy
15	SC 42	ISO/IEC TR 20547-5:2018 - Information technology - Big data reference architecture - Part 5: Standards roadmap

2.4 Data Quality

The following standards, referenced in 4.2.6, focus on the quality aspects of the Data used in ML models.

No.	Publisher	Publication
16	SC 42	ISO/IEC 5259-1:2024 - Artificial intelligence - Data quality for analytics and machine learning (ML) - Part 1: Overview, terminology, and examples
17	SC 42	ISO/IEC 5259-2:2024 - Artificial intelligence - Data quality for analytics and machine learning (ML) - Part 2: Data quality measures
18	SC 42	ISO/IEC 5259-3:2024 - Artificial intelligence - Data quality for analytics and machine learning (ML) - Part 3: Data quality management requirements and guidelines
19	SC 42	ISO/IEC 5259-4:2024 - Artificial intelligence - Data quality for analytics and machine learning (ML) - Part 4: Data quality process framework
20	SC 42	ISO/IEC 5259-5:2025 - Artificial intelligence - Data quality for analytics and machine learning (ML) - Part 5: Data quality governance framework
21	SAE	J3298:2024 - Artificial Intelligence Data for Ground Vehicle Applications

2.5 Testing and Assessment

The following standards, referenced in 4.3, provide guidance on testing of AI systems and evaluation of performance and robustness.

No.	Publisher	Publication
22	SC 7	ISO/IEC TR 29119-11:2020 - Software and systems engineering - Software testing - Part 11: Guidelines on the testing of AI-based systems
23	IEEE	IEEE 2841:2024 - IEEE Recommended Practice for Framework and Process for Deep Learning Evaluation
24	SC 42	ISO/IEC TR 24029-1:2021 - Artificial Intelligence (AI) - Assessment of the robustness of neural networks - Part 1: Overview
25	SC 42	ISO/IEC TR 24029-2:2021 - Artificial intelligence (AI) - Assessment of the robustness of neural networks - Part 2: Methodology for the use of formal methods
26	SC 42	ISO/IEC TS 4213:2022 - Information technology - Artificial intelligence - Assessment of machine learning classification performance
27	IEEE	IEEE 2937:2022 - IEEE Standard for Performance Benchmarking for Artificial Intelligence Server Systems
28	IEEE	IEEE 3129:2023 - IEEE Standard for Robustness Testing and Evaluation of Artificial Intelligence (AI)-based Image Recognition Service
29	IEEE	IEEE 3168:2024 - IEEE Standard for Robustness Evaluation Test Methods for a Natural Language Processing Service That Uses Machine Learning

2.6 Computing Devices

This standard, referenced in 4.4, provides an overview of machine learning (ML) computing devices, including their concepts, characteristics, and approaches for performance optimization.

No.	Publisher	Publication
30	SC 42	ISO/IEC TR 17903:2024 - Information technology - Artificial intelligence - Overview of machine learning computing devices

2.7 Standardization

The following documents, referenced in 5.1, are helpful in understanding the standardization challenge for AI systems employed in ground vehicles.

No.	Publisher	Publication
31	SAE	SAE Edge Report - Navigating the Evolving Landscape of Safety Standards for Machine Learning-based Road Vehicle Functions
32	SAE	SAE J3312 - Artificial Intelligence Use Cases for Ground Vehicle Applications

2.8 Compatibility Challenges with V-Model

These documents, referenced in 5.2, provide guidance on establishing, implementing, and continually improving an AI management system.

No.	Publisher	Publication
33	SC 42	ISO/IEC 5338:2023 - Information technology - Artificial intelligence - AI system life cycle processes
34	SC 7	ISO/IEC/IEEE 12207:2017 - Systems and software engineering - Software life cycle processes
35	SC 7	ISO/IEC/IEEE 15288:2023 - Systems and software engineering - System life cycle processes
36	SC 42	ISO/IEC 8183:2023 - Information technology - Artificial intelligence - Data life cycle framework
37	SC 42	ISO/IEC 42001:2023 - Information technology - Artificial intelligence - Management system
38	SC 42	ISO/IEC 38507:2022 - Information technology - Governance of IT - Governance implications of the use of artificial intelligence by organizations

2.9 Risk Management

The documents listed below, referenced in 5.3, cover topics related to risk management in general and to safety, security, and privacy of AI systems in particular.

No.	Publisher	Publication
39	NIST	AI Risk Management Framework (https://www.nist.gov/itl/ai-risk-management-framework)
40	NIST	AI RMF Playbook (https://airc.nist.gov/AI_RM_F_Knowledge_Base/Playbook)
41	SC 42	ISO/IEC 23894:2023 - Information technology - Artificial intelligence - Guidance on risk management
42	SC 42	ISO/IEC TR 27563:2023 - Security and privacy in artificial intelligence use cases - best practices
43	IEEE	IEEE 3156:2024 - IEEE Standard for Requirements of Privacy-Preserving Computation Integrated Platforms
44	SC 32	ISO/SAE 21434:2021 - Road vehicles - Cybersecurity engineering
45	SC 32	ISO/PAS 5112:2022 - Road vehicles - Guidelines for auditing cybersecurity engineering
46	SC 32	ISO 26262:2018 - Road vehicles - Functional safety, 12 parts
47	SC 32	ISO 21448:2022 - Road Vehicles - Safety of the Intended Functionality
48	SC 32	ISO PAS 8800:2024 - Road vehicles - Safety and artificial intelligence
49	SC 32	ISO TS 5083 - Road vehicles - Safety for automated driving systems - Design, verification, and validation
50	SC 42	ISO/IEC TR 5469:2024 - Artificial intelligence - Functional safety and AI systems
51	SC 32	ISO/PAS 8926:2024 - Road vehicles - Functional safety - Use of pre-existing software architectural elements

2.10 Trustworthiness

The following documents, referenced in 5.6, cover topics such as ethics, bias, controllability, and explainability.

No.	Publisher	Publication
52	SC 42	ISO/IEC TR 24028:2020 - Information technology - Artificial intelligence - Overview of trustworthiness in artificial intelligence
53	SC 42	ISO/IEC TR 24368:2022 - Information technology - Artificial intelligence - Overview of ethical and societal concerns
54	IEEE	IEEE 2894:2024 - IEEE Guide for an Architectural Framework for Explainable Artificial Intelligence
55	SC 42	ISO/IEC TR 24027:2021 - Information technology - Artificial - intelligence (AI) - Bias in AI systems and AI aided decision making)
56	SC 42	ISO/IEC TS 12791:2024 - Information technology - Artificial intelligence - Treatment of unwanted bias in classification and regression machine learning tasks
57	SC 42	ISO/IEC TS 8200:2024 - Information technology - Artificial intelligence - Controllability of automated artificial intelligence systems

2.11 Regulatory Challenges

The documents below are referenced in Section 6.

No.	Publication
58	OECD AI Principles (https://www.oecd.org/en/topics/artificial-intelligence.html)
59	Proposal for a Draft Resolution with Guidance on Artificial Intelligence in the Context of Road Vehicles (https://unece.org/transport/documents/2024/06/working-documents/secretariat-proposal-draft-resolution-guidance)
60	EU AI Act (https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689)
61	EI AI Act Risk-Based Approach (https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai)
62	European Commission: European Artificial Intelligence Act Comes into Force (https://ec.europa.eu/commission/presscorner/detail/en/ip_24_4123)
63	Amendment to EU Regulation 2019/2144 (https://artificialintelligenceact.eu/article/109/)
64	Colorado AI Act (https://leg.colorado.gov/bills/sb24-205#:~:text=On%20and%20after%20February%201,in%20the%20high%20risk%20system.)
65	Japan AI Guidelines for Business Ver1.0 (https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20240419_9.pdf)
66	Canada AI and Data Act (https://www.parl.ca/legisinfo/en/bill/44-1/c-27)
67	UK Automated Vehicles Act (https://www.legislation.gov.uk/ukpga/2024/10/contents)

3. AI SYSTEMS

3.1 Main AI Techniques Classifications

Artificial Intelligence (AI) refers to software and system functionality that exhibits such traits as pattern recognition and capacity to learn, understand, reason, plan, cognate, and problem solve. It refers to the ability of machines to perform tasks that are normally handled by human intelligence.

It is a widespread practice to classify the AI techniques in three main classes: Symbolic AI, Numerical Analysis, and Machine Learning (ML).

- **Symbolic AI:** Solutions are based on human-readable representations of problems, where real-world objects, along with their characteristics, relationships, and interactions, are represented by symbols. A vehicle, for example, could be represented by the textual string “vehicle.”
- **Numerical Analysis:** This technique includes several common algorithmic approaches based on non-data driven techniques used in a wide range of applications including Optimization, Mapping, and State Estimation in applications such as robotics.

Although these techniques do not use model training, massive amounts of data, and “guesswork,” their main limitation lies in the difficulty of codifying to adequately represent the real world, with all its complexity and nuances. Currently, these techniques are not common in practical ground vehicle applications.

ISO/IEC 22989:2022^[1] provides a comprehensive set of terms and definitions related to the following domains: **AI Terms** (AI agent, AI system, automation, cognitive computing, and expert systems); **Data-related Terms** (data annotation, ground truth, imputation, and various data types); **Machine Learning Terms** (algorithms, models, parameters, and training approaches); **Neural Network Terms** (activation functions, network types, and architectures); and **Trustworthiness Terms** (addressing accountability, bias, reliability, and validation).

In this report, following the terminology established in ISO/IEC 22989:2022, an **AI System** refers to an “engineered system that generates outputs such as content, forecasts, recommendations, or decisions for a given set of human-defined objectives.” An AI system consists of AI components such as AI preprocessing component, AI post-processing component, AI model, and conventional software component inside an AI system, etc.

ISO/IEC 22989:2022 also outlines several other AI concepts including the distinction between narrow AI and general AI, agent paradigms, knowledge representation, and symbolic versus sub-symbolic approaches.

- **Machine Learning (ML)** is a subfield of AI that allows machines to learn from data without being explicitly programmed. It is a process by which an AI system derives its knowledge by learning from a training dataset to produce an AI model. The **ML Model** is defined as a mathematical construct that generates an inference or prediction based on input data or information and includes a functionality that is created by machine learning.

Only Machine Learning applications in ground vehicles will be discussed in detail in this document. **Neural Networks (NN)** algorithms are commonly used in ground vehicle applications and as such are outlined further here. Neural networks get inspiration from how neurons are connected in the brains of humans and animals. The structure of neural networks is composed of interconnected processing elements (or “neurons”). Each neuron receives several inputs and generates only one output. Each connection has an assigned weight related to the importance of the input. The neural network “learns” by training with known inputs and comparing actual output with the expected one and using the error to adjust weights. In the process of learning, the links that produce correct answers are strengthened and those that generate incorrect answers are weakened.

Neural networks are organized into layers, where the output of one layer becomes the input to the next layer. It is possible to have hidden layers between the input and output layer. **Deep Neural Network (DNN)**, or “Deep Learning,” specifically refers to neural networks with many hidden layers. Deep Learning can reduce or eliminate the need for feature engineering as the most relevant features are identified automatically but typically use significant time and computing resources.

More information on Machine Learning (ML) framework and generic system components and their functions in the AI ecosystem can be found in both ISO/IEC 24372:2021^[2] and ISO/IEC 23053:2022^[3]. These two standards expand on the Machine Learning algorithms and approaches mentioned in in ISO/IEC TR 24030:2021^[4]. These standards present a comprehensive overview of ML systems with their components: tasks, models, data, and tools. They also summarize various machine learning approaches such as Supervised Learning, Unsupervised Learning, Semi-supervised Learning, Reinforcement Learning, Transfer Learning, and Continuous Learning and provide examples of ML algorithms such as neural networks, Bayesian networks, decision trees, and support vector machines. The ISO/IEC 23053:2022 standard includes an informative annex demonstrating how to describe data flows and develop data use statements for ML processes.

3.2 Difference Between AI and Conventional Methods

Artificial Intelligence (AI), especially those based on machine learning, fundamentally differ from conventional methods in how problems are approached and solved. While conventional methods rely on predefined rules, equations, and logic defined by experts, AI systems learn patterns and relationships directly from data through training. Additionally, AI often uses end-to-end learning, processing raw inputs to generate outputs, unlike conventional approaches that involve modular, step-by-step processes. These distinctions make AI powerful for unstructured problems but also introduce challenges that are not normally associated with the conventional methods.

The core difference between AI and conventional methods are summarized in the following sections.

3.2.1 Inductive versus Deductive

Inductive method involves deriving general principles or patterns from specific observations or data. AI systems, especially those based on machine learning, operate fundamentally through induction: **Learning from Data** (AI models are trained on datasets to discover patterns, correlations, and trends. For instance, an image classification model learns to distinguish cats from dogs by analyzing labeled images.); **Generalization** (Once trained, the AI model generalizes its learning to make predictions on new, unseen data. The system assumes that the patterns observed in the training data will hold true for future data.); and **Uncertainty and Probabilities** (AI systems often express their conclusions probabilistically because inductive reasoning does not guarantee absolute certainty. For example, a model might predict a 90% likelihood that an image contains a cat.).

Deductive methods start with a general principle or rule and applies it to specific cases to derive conclusions. Conventional engineering and rule-based systems often rely on deduction: **Rule-Based Systems** (These systems use explicitly defined rules [e.g., "If X, then Y"] to make decisions. The rules are predetermined based on known laws of physics, logic, or domain knowledge.); **Deterministic Outcomes** (Deductive methods provide deterministic results, assuming the initial premises are correct. There is no probabilistic interpretation unless explicitly built in.); and **Top-Down Approach** (Deduction works in a top-down manner, moving from the general to the specific.).

Table 1 shows the main difference between Inductive AI methods and Deductive conventional methods.

Table 1 - Inductive AI versus deductive conventional methods

Aspect	Inductive (AI)	Deductive (Conventional)
Starting Point	Observations/Data	General principles/rules
Process	Generalizing from specific examples	Applying rules to specific instances
Outcome	Probabilistic, pattern-based	Deterministic, rule-based
Complexity	Handles complex, poorly understood problems	Works well for well-understood, structured problems
Certainty	Uncertain, relies on many things such as data quality and coverage	Certain, relies on correctness of premises

3.2.2 Data-Driven versus Rule-Based

AI, particularly machine learning and deep learning, relies on data to learn patterns, make decisions, and solve problems. Unlike conventional systems, where logic and rules are explicitly programmed, AI learns from examples in the data to infer relationships and behavior. This makes AI especially effective for complex, unstructured, or high-dimensional problems where explicit rules are hard to define.

The key difference of data-driven AI and rule-based conventional methods are summarized in Table 2.

Table 2 - Data-driven AI versus rule-based conventional methods

Aspect	Data-Driven Methods (AI)	Rule-Based Methods (Conventional)
Approach	Learns automatically from data through training	Manual programming, relies on explicitly defined rules, logic, or equations
Dependency	Large and high-quality datasets	Domain expertise to create rules
Complexity Handling	Excels at unstructured, complex, high-dimensional problems	Effective for structured, well-defined problems
Feature Extraction	Automatically extract relevant features from raw data	Features are manually selected, designed, or engineered
Adaptability	Typically involves retraining on new data	Typically involves reprogramming to adapt to changes
Interpretability	Often an opaque “black box” - hard to interpret	Highly interpretable and transparent due to explicit rules

3.2.3 End-to-End versus Modular

It is possible to employ end-to-end methods that process raw inputs directly and generate desired outputs, often bypassing intermediate manual steps. It often involves training and optimizing the entire system as a whole, adjusting all internal components simultaneously. This contrasts with conventional systems, which break the process into distinct, well-defined stages of modules, and step-by-step processes, each handling a specific subtask.

The key difference between end-to-end AI method and conventional modular methods are summarized in Table 3.

Table 3 - End-to-end AI versus modular conventional

Aspect	End-to-End AI Method	Conventional Modular Methods
Architecture	Single component	Multiple distinct components
Optimization	The entire system is optimized as one	Each component is optimized separately
Data Handling	Processes raw input directly to produce output	Intermediate outputs passed between components
Flexibility	Learns tasks holistically from data	Relies on domain knowledge for component design
Error Diagnosis	Hard to pinpoint the source of errors	Easier to identify which component is failing
Interpretability	Often a “black box” with less transparency	More transparent, with interpretable components
Adaptability	Adapts to new data or scenarios by retraining on new data in general without reprogramming	Uses manual modification of components and redesigning and reprogramming for new conditions
Dependency on Data	Relies heavily on large and diverse datasets to learn	Relies more on expert knowledge, predefined rules, equation, or logic than data
Performance	Excels in complex, unstructured tasks	Works well for structured, well-defined tasks

ISO/IEC 5339:2024^[5] provides guidance for the development and application of artificial intelligence (AI) systems, focusing on stakeholder roles, characteristics, and considerations throughout the AI application life cycle. This document differentiates between Functional Characteristics, which include the ability to acquire information, make timely predictions, and improve through continuous learning, and Non-functional Characteristics, which emphasize trustworthiness, including robustness, reliability, resilience, controllability, explainability, predictability, and transparency. In addition, based on their roles and objectives, ISO/IEC 5339 differentiates between three types (Perspectives) of Stakeholders: **Make Perspective** (involves AI producers, developers, and data providers who create the AI application), **Use Perspective** (encompasses AI customers and users who utilize the application for decision-making), and **Impact Perspective** (considers the community and regulators assessing the societal effects and compliance of the AI application).

An interesting concept where an engineering discipline, normally associated with systems based on conventional methods, has found renewed interest in AI systems is Knowledge Engineering (KE). The reference architecture described in ISO/IEC 5392:2024^[6] provides a framework for acquiring, modeling, storing, and applying knowledge to enable intelligent, knowledge-driven services. An AI system with KE can acquire knowledge directly from the information extracted from the data and further construct the knowledge base.

In the constructed knowledge base of the AI system, there are two types of knowledge: **Knowledge of Methods** (ML models that are data-driven and based on subject expertise captured from an expert) and **Knowledge of Contents** (knowledge about subject area in the form of concepts, relationships, entities drawn from texts, videos, etc., which is acquired from the input data or information and from first principles like physics-based models).

ISO/IEC 5392 emphasizes two views similar to the views in the Big Data Reference Architecture (see 4.2.1) to describe AI systems with KE. The **User View** describes the ecosystem of KE with the KE activities, the roles, and the sub-roles. Some of the roles listed are data suppliers, algorithm suppliers, knowledge service providers, knowledge appliers, and knowledge ecosystem partners. The **Functional View** describes the architectural layers and the classes of functional components within those layers that implement the activities of the roles and sub-roles within the user view. Examples are KE Infrastructure layer, KE Construction layer, KE Platform layer, KE Application layer, and multi-layer functions.

Although ISO/IEC 5392 does not provide ground vehicle-specific examples for applications of AI Systems with KE, the following examples can be inferred from the typical applications listed in Annex C: **Expert Systems** (KE can be used to develop an AI-based intelligent maintenance system that can predict component failures and suggest preventive actions), **Decision-Making Systems** (KE-based systems can analyze real-time data from sensors and make decisions to optimize vehicle performance or ensure safety such as in a collision avoidance system. However, it can be a design challenge to ensure that KE systems can process and compute knowledge quickly enough to meet real-time demands.), and **Automatic Content Generation** (KE can generate user-friendly content maintenance instructions or driving tips, such as alerts for tire pressure or oil changes with detailed instructions, based on the vehicle's condition and user behavior.).

4. TECHNICAL CHALLENGES

4.1 ML Models Construction

AI's data-driven, end-to-end approach stems from its ability to learn directly from raw data and optimize the entire process holistically. This makes AI well suited for complex, dynamic, and high-dimensional problems that are difficult or impossible to solve with traditional, modular, or rule-based methods. However, it comes with unique challenges that are often underestimated.

The "black-box" nature of AI arises from its reliance on complex, data-driven, end-to-end, and often opaque models. While this enables AI to excel at tasks that are too complex for conventional methods, it also presents significant challenges in terms of trust, accountability, and safety.

The black box nature of AI makes it unpredictable because it is often challenging to understand, explain, or foresee how an AI model arrives at its decisions or outputs. This unpredictability arises from several inherent characteristics of modern AI systems, particularly those based on complex machine learning models like deep learning.

The challenges of unpredictable model behavior of AI can be observed in a number of scenarios.

4.1.1 Different Training Data Split

When splitting the dataset, the specific instances that end up in the training set may vary, leading to the model being exposed to slightly different patterns. Each split might expose the model to a different subset of features or class distributions. Since the model learns from the training data, even minor differences in the training set (e.g., missing certain outliers or rare cases) can lead to different learned parameters (weights, decision boundaries), resulting in a model that behaves very differently on new data. The impact of dataset splitting can be illustrated from several aspects:

- **Variation in Class Distribution:** Different split can lead to variations in class distributions between training and test sets, especially if the dataset is imbalanced (i.e., some classes are more frequent than others). One split might result in more balanced training data, while another might lead to an imbalance where one class dominates the training set. Variations in class distributions can heavily influence how the model learns to classify various categories. If a split results in some classes being underrepresented, the model may not learn those classes well, and a small shift in data distribution can significantly change how the model performs across various categories.
- **Outliers:** Outliers are data points that differ significantly from the majority of the data. Different split of the data can result in some splits where outliers are concentrated in the training set, while other splits might put those outliers in the test set. When outliers are included in the training set, the model learns to handle them. But if they are in the test set and absent in training, the model may not generalize well to these unusual cases. This can lead to major differences in model performance and make the model overly sensitive to even minor changes in input because outliers tend to shift decision boundaries and influence how the model weights certain features.

- **Shift in Feature Distribution:** Different splits can cause slight shifts in feature distributions between the training and test sets. Even though the overall data distribution might look similar, small variations in the distributions of specific features can affect the model's learning process, leading to different internal representations. Models, particularly complex ones like deep neural networks, are often overly sensitive to the distribution of input features. If a key feature has a slightly different distribution in the training set, the model might learn to interpret it differently. This can cause major differences in how the model makes predictions when it encounters similar data in real-world scenarios or in the test set.
- **Overfitting:** Overfitting occurs when the model learns misleading patterns or correlations specific to the training data rather than capturing generalizable trends. If the training data contains certain patterns or correlations that are not representative of the overall data distribution, the model may be overfit to those specific patterns. Different split of the dataset could lead to scenarios where the training data contains non-representative patterns, causing the model to overfit. This leads to instability, where the model's predictions become overly sensitive to minor changes in the data.

4.1.2 Different Training Parameters

Commonly used parameters, such as learning rate and batch size, are widely recognized for their impact on the overall optimization process during model training and can lead to significantly different models with unique learned weights. However, numerous other factors, parameters, and configurations during training can also substantially affect model weights and performance which lead to different model behaviors.

For example, different optimization algorithms have unique methods for updating weights, potentially leading to different outcomes even when trained on the same data. Adjusting the objective function can alter the focus of the model's learning, resulting in entirely different learned behaviors. When the objective function includes trade-offs (like balancing precision and recall), modifying the importance of these terms can also yield distinct models.

Adjustments to regularization parameters can also significantly affect model weights and performance and can lead to simplifying of the model structure. For instance, L1 and L2 Regularization methods are based on penalizing the model's high-value coefficients and are often applied when the dataset has a large number of features. L1 regularization (also known as Lasso for "least absolute shrinkage and selection operator") adds an absolute value of the sum of coefficients (or the "absolute value of magnitude" of the coefficient) as a penalty term to the model's sum of squared errors (SSE) loss function. L2 Regularization, on the other hand, adds the squared sum of coefficients (or the "squared magnitude" of the coefficient) as a penalty term to the model's SSE loss function. It is possible to reduce coefficient values to zero using L1 Regularization but not by using L2 regularization, which can reduce coefficient values toward zero but never exactly to zero.

It is essential to note that many additional factors, often overlooked or inaccessible to the typical user, can also influence model outcomes. For instance, initial weight values can lead a model to converge at different local minima, and using different random seeds for data shuffling can create varied training batches, affecting the model's learning path over time. This inherent randomness means that models may learn different weights in separate training runs, which profoundly impacts final weights and overall performance.

4.1.3 Instance Level Performance

Objective functions (also referred to as the loss function or cost function) stand at the core of AI/ML. An objective function provides a formal, mathematical specification of the problem and quantifies how well a model's output aligns with the target values in the training data. The objective function is critical because it defines what "good" performance means for the model and serves as the guide during the training process to update the model's parameters.

Most objective functions in AI/ML measure performance at the dataset level, focusing on aggregate accuracy or another metric over the entire dataset rather than individual predictions. For example, the model is trained to minimize the overall error (e.g., sum of squared errors or cross-entropy) across all training instances. This can lead to two models having similar performance on a dataset level but different predictions for specific individual instances.

Since the objective function focuses on the overall dataset performance, there is flexibility for the model to arrive at different predictions for individual data points as long as the aggregate performance metric remains similar.

This means that two models with the same accuracy or loss value might make very different predictions on certain instances within the dataset. These differences at the instance level can have significant implications, especially in safety-critical applications where individual predictions are crucial.

4.1.4 Model Compression and Pruning

Model pruning and compression are often neglected when discussing model behaviors during model training. These techniques aim to make models more efficient while maintaining similar performance levels. However, due to the sensitive nature of AI/ML models, pruning and compression can change the model behaviors, where minor changes lead to significant impacts on the model's weights, performance, and behavior.

Model compression involves techniques that reduce the size of the model. Popular methods include quantization (reducing the precision of weights) and knowledge distillation (training a smaller model to mimic a larger one).

In quantization, the precision of the weights is reduced (e.g., from 32-bit floating point to 8-bit integers). While this saves memory and speeds up computation, reducing precision can introduce rounding errors and alter the values of the model's learned parameters. These minor differences can lead to changes in how the model processes input data. For example, reducing the precision of weights may cause the model to misclassify inputs that are close to the decision boundary.

In knowledge distillation, a smaller model (the “student”) is trained to replicate the behavior of a larger model (the “teacher”). Although the student model captures much of the teacher's knowledge, its architecture and weights differ significantly from those of the larger model. This reduced capacity can lead to performance differences, especially on edge cases that the larger model manages well but the student model may struggle with due to its reduced capability.

Model pruning is the process of removing certain weights, neurons, or even entire layers or branches from a model that are deemed to have minor impact on the final output. The goal is to reduce the model's complexity without significantly affecting its performance. Pruning typically targets weights that have a low magnitude, under the assumption that these weights contribute less to the model's predictions. However, the removal of even seemingly insignificant weights can change the internal decision-making process of the model which causes the behavior of the model changes. Moreover, by altering the structure of the model (i.e., removing connections or neurons), pruning changes the learned decision boundaries, thereby affecting overall model behavior.

4.1.5 Model Fine-Tuning and Update

After the AI/ML model is deployed and is operational, maintenance and update are in general required to sustain the capability of the system to provide services. This often involves collecting data for corner cases that the model does not perform well and occasionally fine-tuning or updating the model to improve the model performance in the deployment environment.

Fine-tuning a model to handle corner cases, while intended to improve performance for rare or extreme situations, leads to making minor, targeted adjustments to handle specific scenarios. This, in turn, can lead to broad and unintended changes in the model's overall behavior, illustrating the “changing anything changes everything” effect.

Fine-tuning adjusts the model's weights to improve performance on the new, corner case data. However, the model's weights are interconnected and optimized for the entire training set, so changing them for a small portion of new data can totally change the weights of model. Adjusting the weights can lead to significant changes in how the model behaves across the broader input space, not just in the corner cases. Slight changes in weight values can have a cascading effect across the entire model, influencing how the model interprets both new and old data.

Furthermore, fine-tuning for corner cases changes the model's decision boundaries, the surfaces that separate different classes or predictions in the feature space. The new data points from the corner cases push these boundaries, potentially moving them away from where they were initially set for common cases. A decision boundary shift caused by corner case data could lead to incorrect classifications for data points that were previously classified correctly. The entire feature space can be subtly altered, and inputs that used to lie comfortably within one class might now be misclassified due to the boundary adjustment.